



Humans 40,000 y ago developed a system of conventional signs

Christian Bentz^{a,b,1} and Ewa Dutkiewicz^c

Edited by Melinda Zeder, Smithsonian National Museum of Natural History, Washington, DC; received July 28, 2025; accepted January 14, 2026

As humans, we store and share information. This allows us to distribute knowledge necessary for survival and to coordinate large groups. Our hominin ancestors harnessed the surfaces of mobile artifacts and cave walls as information carriers since the Paleolithic time period. Theories abound as to the meaning and function of these Paleolithic signs. However, very little is known about their basic, measurable properties. We here analyze a corpus of more than 200 mobile objects of a 43,000 to 34,000 y old Aurignacian culture—associated with the first modern humans to settle in Central Europe. These objects are adorned with several thousand geometric signs. We apply classification algorithms and statistical models to capture their quantitative properties. First, our analyses illustrate that these sign sequences are clearly distinguishable from modern day writing. Second, however, their statistical properties are comparable to sign sequences on the earliest protocuneiform tablets. Third, Paleolithic signs were systematically applied to yield higher information density on certain types of objects, e.g. ivory figurines compared to tools. These results cannot be taken to strictly prove that Aurignacian sign sequences encoded numero-ideographic information as in the case of protocuneiform. However, they prove that the first hunter-gatherers arriving in Europe already applied sign sequences of comparable complexity in a deliberate, systematic, and conventional manner—several ten thousand years before the advent of genuine writing.

Paleolithic signs | human evolution | quantitative linguistics | protocuneiform | information theory

As humans, we store information outside of our minds. This capacity for information encoding beyond the here and now of spoken words is nowadays reflected in symbols and written language. It is a hallmark of human cognitive evolution and underlies artificial computing systems. Without it, the information age would not be possible. Archaeologists have identified reflections of this capacity in excavation layers dating back to the Paleolithic period of 400,000 to 15,000 thousand years ago (1–5).

At the beginning of the Upper Paleolithic around 45,000 calibrated years before present (cal BP), modern humans arrived in Eastern and Central Europe (6, 7). On their journey they encountered their distant relatives—Neanderthals (8–11). In this time of migrations and population turn-overs, modern humans produced a panoply of so-called mobile objects, such as tools and figurines made of ivory, bone, or antler. These are present right from the earliest period of the Upper Paleolithic, the so-called Aurignacian technocomplex. Especially the Dordogne region in southwestern France (12–14), the cave systems of the Swabian Jura in southwestern Germany (15–17), and a cluster of sites in Belgium (18, 19) have yielded hundreds of objects adorned with sequences of geometric signs.

In earlier studies, selected artifacts were analyzed microscopically and experimentally to demonstrate that so-called artificial memory systems (AMS) might have their beginnings already in the late Middle Paleolithic associated with Neanderthals of 70,000 to 40,000 y ago (3, 20, 21). Namely, a few mobile artifacts (e.g. raven and hyena bones) bear regular incisions potentially encoding numerical information. Some traces of artificial memory systems or “exograms” might even reach back to the Lower Paleolithic several hundred thousand years ago (22, 23).

Having said this, there is a clear increase in the number of mobile artifacts with geometric markings in assemblages of the Upper Paleolithic period (5, 12–14, 16, 24). Moreover, the codes of artificial memory systems seem to have become more complex toward the end of this period (3). How to exactly quantify and model this complexification remains unclear. We here provide openly published data and computational tools to assess the complexification of sign sequences produced by humans over evolutionary time.

In general, our study builds on research applying computational tools to pin down the statistical properties of human language in comparison to ancient sign systems

Significance

Humans have carved visual signs into the surfaces of mobile artifacts and cave walls since several hundred thousand years. We here analyze a 40,000 y old assemblage of mobile artifacts bearing sequences of intentionally engraved geometric signs. These sign sequences have a complexity comparable to the earliest protocuneiform and were selectively applied to yield higher information density on figurines than on tools. This proves that the first hunter-gatherers arriving in Europe already developed a system of intentional and conventional signs on mobile artifacts. Our study more broadly relates to research into statistical properties of human language and writing compared to other sign systems.

Author affiliations: ^aDepartment of Language Science and Technology, Saarland University, Saarbrücken 66123, Germany; ^bFaculty of Humanities and Cultural Studies, University of Passau, Passau 94032, Germany; and ^cStone Age Department, Museum für Vor- und Frühgeschichte, Staatliche Museen zu Berlin, Berlin 10117, Germany

Author contributions: C.B. and E.D. designed research; C.B. and E.D. performed research; C.B. contributed new reagents/analytic tools; C.B. analyzed data; and C.B. and E.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

PNAS policy is to publish maps as provided by the authors.

¹To whom correspondence may be addressed. Email: chris@christianbentz.de.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2520385123/-/DCSupplemental>.

Published February 23, 2026.

(25–32) as well as animal communication systems (33–36). More specifically, it relates to archaeological investigations of Paleolithic signs on mobile objects (3, 17, 20, 21, 32, 37) as well as cave art (38–42)—sometimes explicitly linked to numerical cognition (43). Also, experimental work emerges to understand how humans perceive the geometric patterns on mobile artifacts and parietal art, and to clarify their semiotic status (44–47). This is embedded more generally into experimental work on the evolution of semiotic systems in the lab (48–50), as well as the cognitive biases shaping them (51, 52). Our study can thus be seen as part of an emerging field of research: Evolutionary Semiotics.

Signs of the Swabian Aurignacian. Our analyses are based on assessment of overall 260 mobile artifacts from the Swabian Aurignacian (17)—a cluster of cave sites in southwestern Germany. The people inhabiting these caves between 43,000 to 34,000 y cal BP (53) have produced a specialized range of tools to cut meat, work animal hides, and create clothes and ropes (54). They have developed the first musical instruments—flutes—made of bones and ivory (55). Moreover, they have left behind symbolic artifacts, such as beads and pendants for personal ornamentation which bear a local signature (56). Finally, the assemblages are enriched by several dozen figurines laboriously carved out of ivory, reflecting the natural environment as well as a spiritual world unknown to us. This includes animal species living in the area at the time such as woolly mammoths (*Mammuthus primigenius*), wild horses (*Equus ferus*), steppe bison (*Bison priscus*), cave bears (*Ursus spelaeus*), cave lions (*Panthera spelaea*), etc. Besides animal depictions, we also find anthropomorphic depictions including female figurines (57), as well as hybrid creatures between human and cave lion (58).

We are interested in the sign sequences carried by these mobile artifacts. As the most general definition of a “sign” in the context of archaeological material we propose: a modification of a surface which can be perceived and interpreted by an intelligent viewer. Note that this includes bite marks, cut marks, utilitarian manipulations, etc. In our analyses, we restrict this further to intentional and nonutilitarian signs. “Intentional” means that they are not the by-product of another activity, e.g. cut marks as a result of butchering. “Nonutilitarian” means there is no indication that these surface modifications had purely practical functions relating to craftsmanship, e.g. holes for attaching ropes, or circular incisions to help with hafting of spear points. The inventory of sign types includes basic shapes such as lines, points, crosses, but also more complex patterns such as stars, grid patterns, and zigzag lines.

The assignment of sign types to UTF-8 characters is based on first-hand microscopic analyses of the second author. Further details on the materials and analyses for the artifacts of the Swabian Aurignacian are published (17). The publication of the first version of the database (5) includes an evaluation of sign type codings given alternative choices by other researchers who are familiar with the material. The agreement scores for alternative sign type assignments are 91 to 94%, and Cohen’s Kappa scores are 0.29 to 0.44. Further details and examples are given in *Materials and Methods*. Sequences of signs are found on the surfaces of different types of mobile artifacts (Fig. 1), and they are particularly pervasive in four cave sites of the Swabian Aurignacian which are within hiking distance of one another (SI Appendix, Fig. S1).

As comparative samples we employ material from the earliest periods of protocuneiform in ancient Mesopotamia: Uruk V

(approx. 3500–3350 BC), Uruk IV (approx. 3350–3200 BC), and Uruk III (approx. 3200–3000 BC). See Fig. 2 for example tablets. We also harness the Text Data Diversity (TeDDi) sample of 89 different languages written in 16 scripts (59) as a modern day counterpart.

Based on these samples we, first, estimate statistical features from quantitative linguistics for all the approx. 3000 sign sequences in these corpora: type-token-ratios, unigram entropies, entropy rates, and repetition rates. See SI Appendix, Fig. S2 for an overview of the workflow. Second, we input the feature values to classification algorithms. This enables us to assess how similar or different the sequences are given their labels, i.e. Aurignacian, protocuneiform, and modern day writing. Third, we fit multiple regression models for the Aurignacian sequences to predict their information density given meta-information, i.e. the type of artifact (tool, figurine, personal ornament, and others), the volume of an artifact, its preservation, its age, etc. For further details see *Materials and Methods*.

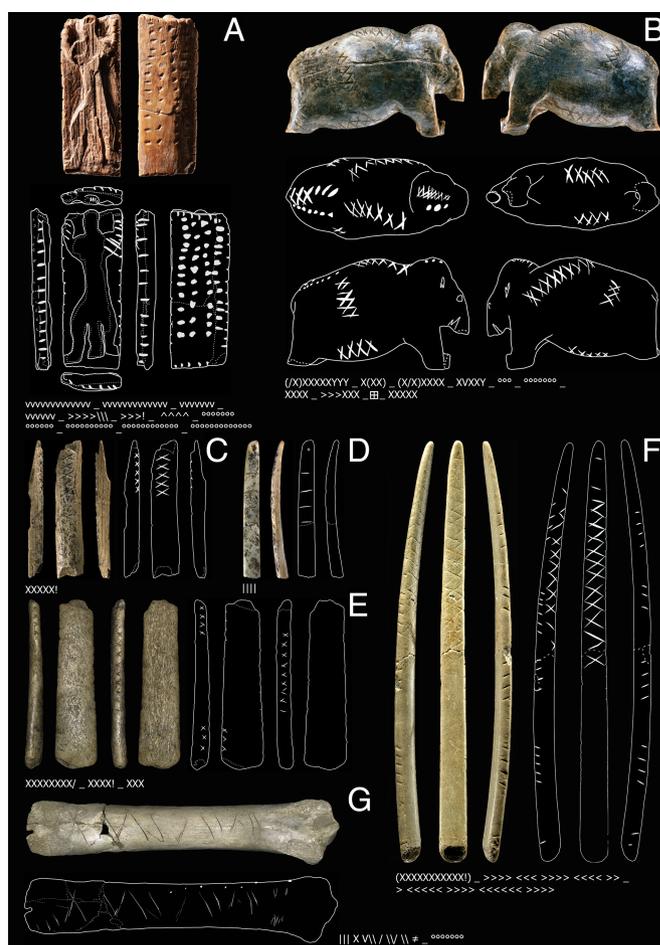


Fig. 1. Mobile artifacts with geometric signs of the Swabian Aurignacian. (A) Plaquette with hybrid creature (so-called “Adorant”), ivory, Geissenklösterle (gkl0025), Landesmuseum Württemberg, Hendrik Zweitasch. (B) Mammoth figurine, ivory, Vogelherd (vhc0145), University of Tübingen, Juraj Lipták. (C) Rod/bâton, ivory, Vogelherd (vhc0001), University of Tübingen, Ewa Dutkiewicz. (D) Personal ornament, ivory, Geissenklösterle (gkl0006), University of Tübingen, Ewa Dutkiewicz. (E) Spatula/lissoir, bone, Vogelherd (vhc0017), University of Tübingen, Ewa Dutkiewicz. (F) Spatula/lissoir, bone, Vogelherd (vhc0162), University of Tübingen, Juraj Lipták. (G) Undetermined, bone, Hohle Fels (hfc0006), University of Tübingen, Ewa Dutkiewicz. Drawings by Ewa Dutkiewicz. Copyright: CC-BY-SA 4.0. For further details on sign coding and preprocessing see *Materials and Methods*.

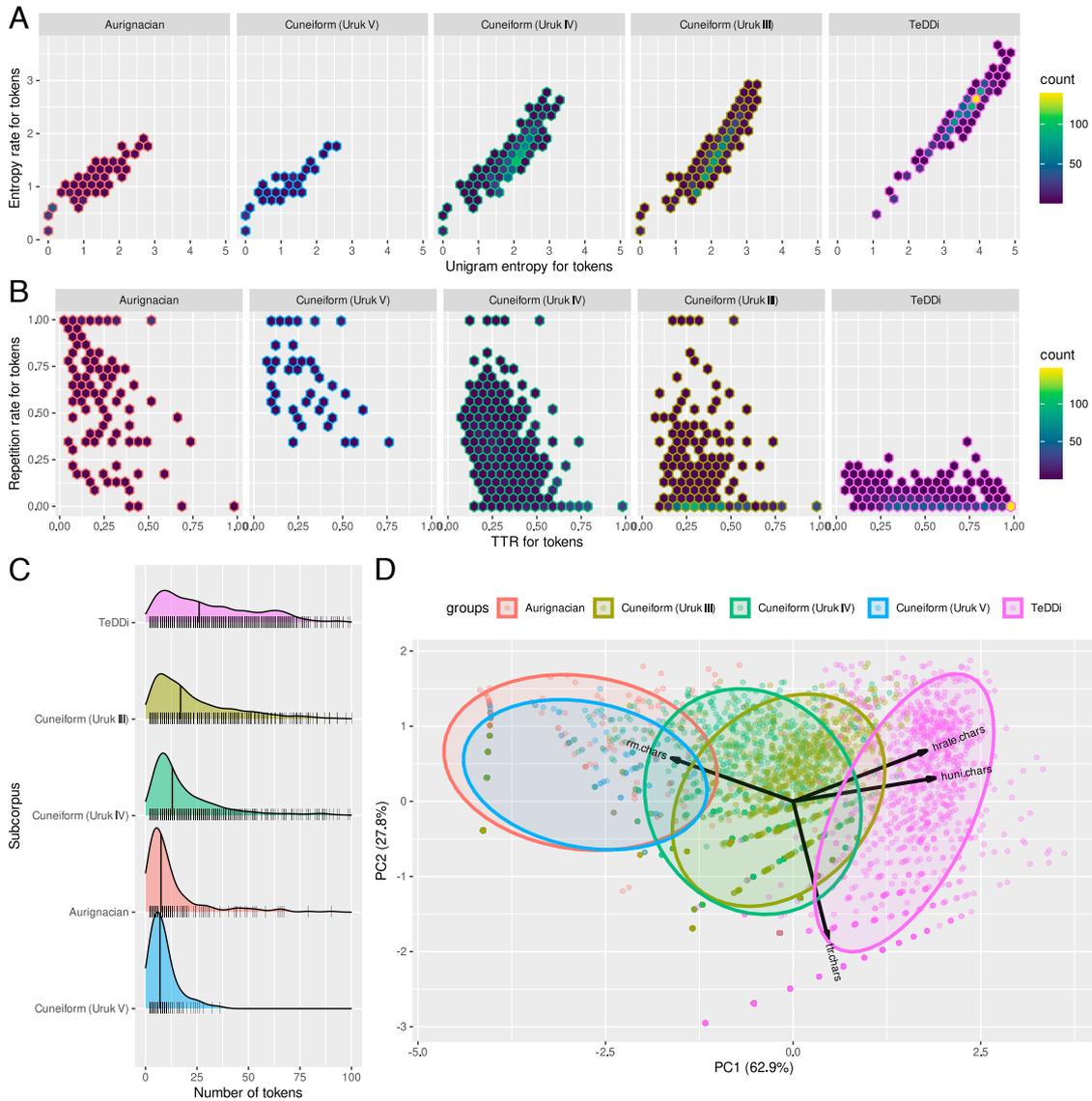


Fig. 3. Statistical feature distributions. (A) Hexagonal bin plots of unigram entropies and entropy rates for Aurignacian, protocuneiform, and modern writing character sequences. The x-axis and y-axis are divided into 20 bins. The color of each bin reflects the data point count in it. (B) Type-token-ratios (TTR) and repetition rates. (C) Distributions of sequence lengths per subcorpus. Black vertical lines give median values. Each vertical tick on the x-axis represents the length of a sequence in UTF-8 characters. (D) Scatterplot of data points for the first two principal components of a PCA on the four statistical features. Normal data ellipses are given colored by the respective group. Arrows represent the correlations of the statistical feature variables with PC1 and PC2.

Discussion

Against the backdrop of our results, we discuss common questions raised about mobile artifacts with markings of the Swabian Aurignacian in particular, and the Upper Paleolithic period more generally.

The Question of Decoration. Recent studies have measured the regularity of notches on bones to determine whether they are more or less visually striking as a decoration (3, 20). Increasing the regularity of distances between notches—up to the differences just about perceivable by humans—is argued to enhance the decorative value. Such technological and experimental analyses are useful to thoroughly understand the production processes behind a given mobile artifact.

On the other hand, categories such as “decoration” and “numerical system,” or “decoration” and “writing system” are not mutually exclusive. Rather, sign systems can be used as decoration

without losing their information value. This is exemplified in historic times by calligraphy, inscriptions on pottery and temples, tattoos of graphemes on human bodies, and many other artistic expressions. “Information density” in an information-theoretic sense is a fundamental property of a sign sequence, irrespective of whether there is a human present to interpret it—or merely find it aesthetically pleasing.

Types of Signs. In the context of semiotics, signs are typically divided into indices, icons, and symbols (60, p. 86). In a nutshell, an index is a sign which stands in a direct causal relation with the object it represents, while an icon represents an object by means of resemblance/similarity. Finally, a symbol is a sign with a conventional and arbitrary (i.e. noniconic) relation to the object.

Some of the signs of the Swabian Aurignacian are likely indices. For instance, notches applied to flutes for indicating where holes need to be cut out. In fact, these straddle upon the category of “utilitarian.” They are still included in our sample since in

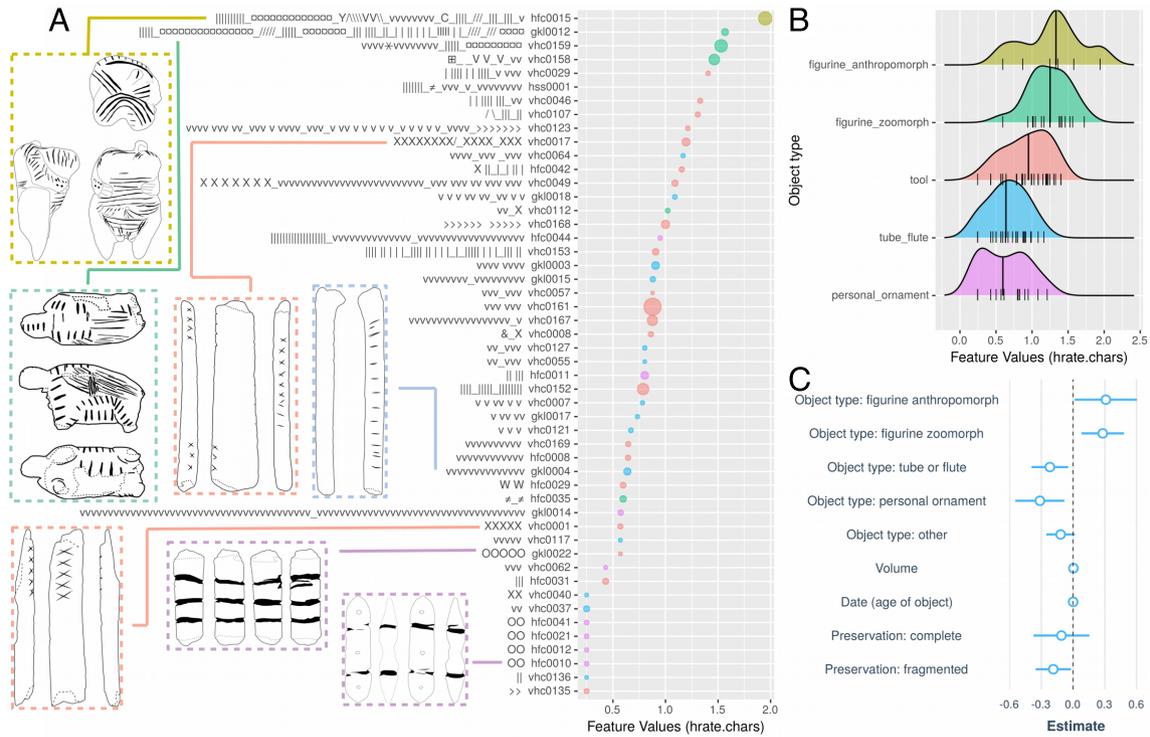


Fig. 4. Results of multiple regression analyses. (A) Sequence plot. Aurignacian sign sequences (50 out of the overall 213 randomly sampled for visual illustration) with their respective entropy rates (x-axis). Colors indicate the type of object carrying a given sequence. The size of dots reflects the volume of the object (in cm^3). (B) Entropy rate distributions per object type. Black vertical lines represent median values. (C) Coefficient plot. Estimated coefficients for a multiple linear regression model predicting entropy rate per sequence of a given object (response variable) from object type (factor variable, reference level “tool”), material (factor variable, reference level “antler”), estimated volume of object (continuous variable), and preservation of object (factor variable, reference level “almost complete”). The dots represent coefficient values, the blue lines are 0.95 CIs. A given coefficient is significantly different from zero if the CI does not intersect with the dashed line.

some cases they occur without any obvious function relating to flute production. There are also credible examples of icons. For instance, on realistically crafted figurines of felines and fish we find dots. These mimic patterns which were discernible by the Paleolithic viewer when observing actual cave lions and trout. Interestingly, however, we also find dots on one side of the ivory plaquette which bears a relief of a hybrid creature between human and lion—the so-called “Adorant”—on the other side (Fig. 14). In this case, the dots seem abstracted away from their original iconic usage, as well as linearly organized in rows of similar lengths, potentially representing numerical information. This would constitute a rather symbolic usage.

The Meaning of Paleolithic Signs. Given the variable nature of meanings associated with signs in human cultures, it is not surprising that there is a multitude of hypotheses about the meanings of Paleolithic signs in general, and the Swabian Jura in particular. For instance, crosses found on mammoth and horse figurines have been interpreted as either simple reflections of fur, counting devices, or symbols of ritual killings (17, pp. 69–80). The latter are found also in association with hunter-gatherer cultures in historic times (61). Marshack (24) claimed that some markings on cave walls and mobile objects encode lunar calendars. Leroi-Gourhan, on the other hand, discusses the hypothesis of so-called “hunting tallies,” as well as similarities to ritual message sticks in the ethnographic record (39, pp. 188–190). The apparent similarity to historical message sticks is an interesting avenue for future research. The recent development of a database for Australian message sticks is a first step in this direction (62, 63).

Linking back to the discussion of early protocuneiform, Sauvet et al. (64, p. 556) identified a set of recurring Paleolithic signs which resemble ideographs in ancient Sumerian, Hittite, Egyptian, and Chinese protowriting. For instance, branch-shaped signs of the Paleolithic resemble ideographs in protowriting denoting “grain” or “herb.” However, “decipherment” of Paleolithic signs in a strict sense is severely hampered by the sheer time depth of this material.

Most recently, it was hypothesized that sign sequences associated with animals in cave art and on mobile objects across the Upper Paleolithic of Europe constitute “phenological calendars” (42). In this interpretation, “Y” signs denote “giving birth,” and sequences of lines or dots represent natural numbers reflecting the passage of time in lunar months. Such calendars would have helped ancient hunter-gatherers to store crucial information about the life cycles and yearly migrations of their prey. Some methodological criticisms of this particular account have been published recently (65–67).

While statistical analyses by themselves cannot conclusively answer questions about meanings of sign sequences, they can, at a minimum, help to narrow down the range of possible interpretations. Given samples of sequences for which we know the interpretation, we can test whether a set of unknown sequences displays a similar statistical fingerprint. The same statistical fingerprint constitutes a necessary but not sufficient condition for functional equivalence.

The Question of Writing. Whenever unknown signs are discovered in archaeological contexts, a natural first question to ask is whether they constitute “writing.” The standard philological

definition of writing refers to the tight link between spoken language, on one hand, and the graphical marks representing it, on the other:

“Broadly defined, writing represents speech. One must be able to recover the spoken word, unambiguously, from a system of visible marks in order for those marks to be considered writing. [...]” (68, p. 18)

Our results strongly contradict the hypothesis that the sign sequences of the Swabian Aurignacian constitute writing in this strict sense. Their statistical properties are very different from those of genuine writing systems around the world. Rates of adjacent sign repetitions are generally very high in Aurignacian sequences, while they are generally very low in modern day writing systems. This is true regardless of whether we consider alphabets, syllabaries, or morpho-syllabaries (“logographies”). While in our sample there are occasional examples of adjacent character or morpheme repetitions in alphabetic writing (French *J’essaie* “I try”), and morpho-syllabic writing (e.g. adjacent repetition of the Hanzi character for the negation particle *bù* in Mandarin Chinese), these are far and few between.

It has been pointed out before that human languages exhibit repetition avoidance at multiple levels of structure (e.g. phonemes, morphemes, words) (69–71). In natural languages, the same structural elements are unlikely repeated in adjacency. This leads to higher information density (more information in a given time interval), potentially linked to cognitive biases of production and usage (71). Against this backdrop, it is very unlikely that the repetitive sign sequences on mobile artifacts represent the structure of languages spoken by the Aurignacian people of the Swabian Jura—the sole criterion for writing *sensu strictu*.

However, it is noteworthy that this strict definition of writing has been repeatedly called into question:

“This would mean that writing is, as some linguists assume, a device for the recording of speech and that all the stages in which writing does not serve this purpose are only feeble attempts in the direction of writing, but not real writing [...] This restriction of the definition of writing is unsatisfactory, however, because it does not take into account the fact that both stages have one identical aim: human intercommunication by means of conventional visible marks.” (72, p. 12)

The Aurignacian sign sequences meet this broader criterion. Namely, they were produced consistently by different individuals over many generations, and they were systematically employed more productively on figurines than on other types of objects (cf. Fig. 4). Also, some sign types have been preferentially used with certain object types (*SI Appendix, Fig. S5*). For example, crosses are some of the most frequently occurring signs. In our sample, however, they never occur on anthropomorph figurines, but rather on zoomorph figurines (especially horses and mammoths) as well as on tools. Inversely, dots occur most frequently on anthropomorph figurines and also zoomorph figurines (especially felines), but never on tools. Note that this cannot be due to material constraints, as both figurines and (some) tools are carved out of ivory, and have surfaces which allow for application of either crosses or dots. Hence, there is a deliberate choice behind adorning humans and felines (cave lions) with dots, but other animals like horse and mammoth as well as tools with crosses. The members of the Swabian Aurignacian culture have certainly handed down such conventions over generations of ivory carvers,

otherwise the occurrence of these statistical patterns over 10,000 y would be extremely unlikely.

However, we do not want to conceal the fact that the human intercommunication criterion is rather loose, and also applies to traffic signs, license plates, Morse code, tattoos, message sticks, and many other devices storing information outside of human minds for intercommunication purposes. When talking about “writing” the strict definition is appropriate. It allows us to further subdivide the vast space of human sign systems into those which are “linguistic” vs. those which are “nonlinguistic” (73). In fact, this division is supported by the empirical evidence, as written languages carry a statistical fingerprint related to structural biases in spoken, and likely also signed languages (29).

In sum, the sign sequences of the Swabian Aurignacian might be described as human intercommunication by means of conventional visible marks, while they certainly do not meet the criterion of writing *sensu strictu*.

Taxonomy of Sign Systems and Writing. Writing in the strict sense—i.e. sign sequences structurally representing spoken language—has developed independently at least three to four times, namely, in Mesopotamia, Egypt, China, and Mesoamerica (68), and potentially a fifth time in the Eastern Pacific (31). These developments fall in the range of roughly 2500 BC to 1500 AD.

However, it has been argued before that the development of writing can only be understood against the backdrop of prehistoric sign systems (74, p. 89). This is underlined, for instance, by analyses of the so-called *La Marche* antler, likely deriving from the Upper Magdalenian period of western France—roughly 15,000 y ago. It is adorned with rows of different sign types which have been interpreted as part of a complex artificial memory system (37).

A possible direction for future research is to craft a taxonomy of sign systems from the Paleolithic to the modern day (73). This requires a set of measurable design features, such as the sign inventory size, linear arrangement, combinatoriality, conventionalization etc. Arguably, some of these design features are already present in the Aurignacian period of 40,000 y ago, e.g. sign inventory size and (to some extent) linear arrangement, while others are largely lacking, e.g. (adjacent) combinatoriality of different sign types, or cannot be strictly proven, e.g. the rebus principle linking sounds of spoken language to signs.

Such a taxonomy will enable a multidimensional comparison between different human sign systems of which writing is but one particular facet.

Commonalities and Differences with Early Protocuneiform.

Our results illustrate that the statistical properties of sign sequences from the Swabian Aurignacian are very similar to those of the Uruk V protocuneiform period (roughly 3500 to 3350 BC). For the later protocuneiform periods of Uruk IV (roughly 3350 to 3200 BC) and Uruk III (roughly 3200 to 3000 BC), however, we already find a significant divergence toward higher information encoding potential.

In this context, it should be noted that the latter periods reflect the transition from mostly numeric to so-called numero-ideographic notations—with a considerable increase in the number of signs in general, and ideographic signs in particular. While the sign list for Uruk V features 47 signs of which 39 (83%) are numeric and only 8 (17%) are ideographic, the list for Uruk III (approximately 150 to 350 y later) features 838 signs, of which only 105 (13%) are numeric, and 733 (87%) are ideographic.

The signs on purely numeric tablets are typically ordered according to their values (75, p. 52). For example, the sign produced by sideways indent of a reed stylus and denoted as N01 might represent the number one, and the sign produced by a vertical indent of a reed stylus and denoted as N14 might represent the number 10. See the Uruk V tablet in Fig. 2A as an example. The numerical values represented by signs are inferred for later periods by interpreting the bundling of signs and the maximum number of repetitions of a given sign. However, note that such bundling and maximum numbers are not always consistent especially in the earliest protocuneiform tablets assigned to Uruk V. Namely, there are examples on tablets where the repetitions of alleged lower number signs transgress the maximum numbers expected (75, p. 51).

The few numero-ideographic protocuneiform tablets assigned to Uruk V typically feature number signs in combination with ideographic signs representing commodities (vessels with some content, sheep, grain, etc.). See as an example the Uruk V tablet in Fig. 2B. This carries the sequence

N14 N14 _ N01 N01 N01 N01 N01 _ DUG~b, [1]

where DUG~b denotes an iconographic representation of a vessel of unknown content. Due to problems of assigning clear numeric values to these early number signs, it is not possible to give an exact translation of this sign sequence. The meaning of this sequence for us is approximately: “some number of vessels of unknown content.”

In comparison, for the material of the Swabian Aurignacian, there are no descendant sign systems which could be used to infer the exact functions of the earlier signs. In the case of particular artifacts like the “Adorant” (Fig. 1A), it has long been argued that these constitute deliberate numerical representations, as the sequences are bundled to similar lengths (10 to 13 dots vertically and 4 horizontally on the flat surface of the plaquette), and these lengths are not simply the outcome of the material constraints (76). In combination with the anthropomorph figurine, this would constitute a numero-ideographic representation. However, in contrast to protocuneiform, there is currently no evidence of higher numbers being encoded with separate, more complex signs.

Protocuneiform in Mesopotamia was exapted to reflect the Sumerian language morpho-phonetically by the Early Dynastic Fara period of approx. 2500 BC (75, p. 80)—i.e. around 1000 y after the earliest period of Uruk V numero-ideographic tablets. This process was fueled by the complexification of the economy (cf. 75, 77). The growing economy required, first, a drastic increase in sign types representing numerals, measures of time, and commodities, and second, an administrative apparatus of scribes. These scribes would gradually transform the signs from reduced depictions to abstract symbols via repetitive usage. This abstraction, in turn, helped to detach the signs from their original referents, and exapt them to represent the sounds of the Sumerian language instead (78).

In stark contrast, we do not detect any significant change in the sign repertoire or in the information density of sequences in the case of Swabian Aurignacian mobile artifacts over the course of 10,000 y.

The Cognitive Implications. Our findings relate to the hypothesis that human intelligence arose through “expanded information capacity” (79). Namely, from an information-theoretic perspective, producing the Aurignacian sequences arguably required roughly as much “information capacity” as the earliest protocuneiform

(Uruk V), but clearly less than later protocuneiform periods (Uruk IV and Uruk III) as well as writing systems around the world. This has to be seen against the backdrop of theoretical results showing that the entropy is an upper bound on the mutual information between signs and referents (80). In other words, the entropy of a sign system is a measurable restriction reflecting the potential to encode sign/referent mappings unambiguously. Our analyses hence suggest that the first hunter-gatherers arriving in Central Europe more than 40,000 y ago already had the information capacity to create a sign system comparable to protocuneiform in terms of information encoding potential.

Conclusions

In conclusion, the people of the Swabian Aurignacian were among the first modern humans to settle in Central Europe. They inhabited the caves of the Lone and Ach Valleys around 43,000 to 34,000 calibrated years ago. They have left behind a panoply of tools and mobile artwork—testimony to their technical skills, and their rich culture. This includes a collection of several dozen ivory figurines, representing animals occurring in the environment at the time, as well as mystical figures of an imaginary world forever lost to us. A subset of these mobile artifacts carry sequences composed of overall more than 3000 signs intentionally carved into their surfaces.

Our analyses show that these sequences are clearly statistically distinct from those generated with modern day writing systems to represent spoken languages. However, they have a very similar “statistical fingerprint” as the earliest numeric and numero-ideographic protocuneiform tablets stemming from the Uruk V period of 3500 to 3350 BC. Moreover, the Aurignacian sequences were not indiscriminately applied to different objects, rather, ivory figurines carry the sequences of highest information density— independent of material constraints such as volume and preservation. The hunter-gatherers of the Swabian Aurignacian have hence developed a sign system with some incipient design features also found in writing, that is, an inventory of different sign types and their linear arrangement, but lacking other design features, e.g. productive combinatoriality of different sign types as well as the rebus principle.

It remains hard—or impossible—to prove that Aurignacian sign systems served the same numero-ideographic functions as protocuneiform. Moreover, there is another stark contrast between them: Protocuneiform developed into a full-blown writing system representing the Sumerian language within the subsequent 1,000 y. The sign sequences of the Swabian Aurignacian, on the other hand, were stable in terms of information density—for 10,000 y—and then disappear.

Materials and Methods

The corpus for the current analyses consists of overall 260 artifacts (213 after preprocessing) carrying signs for the caves of the Swabian Jura. Most such objects derive from Vogelherd Cave in the Lone Valley and Hohle Fels Cave in the Ach Valley. The most common (identifiable) object types carrying signs are tools, zoomorph figurines, tubes and/or flutes, and personal ornaments. The vast majority of these mobile objects is made of mammoth ivory, followed by bone, and antler. An overview of these statistics is given in *SI Appendix, Fig. S6*. Drawings of objects and the identified sign sequences are made available via the SignBase website (<https://www.signbase.org/>) (5). The sign types and sequence codings are also further explained at <https://www.signbase.org/description/> under the point “sign coding.” A table with up-to-date UTF-8 definitions of sign types can be downloaded at <https://www.signbase.org/download/>.

Protocuneiform (Uruk). Information on protocuneiform tablets is taken from the Cuneiform Digital Library Initiative (CDLI) (81). This online database provides pictures, meta-data, transliterations, and standardized sign lists for automated processing. In our analyses, we include transliterations for the Uruk V (roughly 3500 to 3350 BC), Uruk IV (roughly 3350 to 3200 BC), and Uruk III (roughly 3200 to 3000 BC) periods. For some example tablets see Fig. 2. Note that the exact dating of these periods—and the attribution of particular tablets to these—is complicated by the occurrence of so-called secondary deposits (75), and more general problems relating to the organization of excavation campaigns (82). Regardless of their exact dating, however, the respective tablets are commonly discussed as the precursors of cuneiform writing (74, 75, 83, 84). For the earliest period of Uruk V, the number of available transliterations is ca. 100, while for the latter two there are several thousand. In order to not heavily overrepresent the latter in the analyses, we take subsamples of 1,000 for each.

Modern Written Languages. Modern day languages are represented by the Text Data Diversity sample (TeDDi) (59). It features 89 languages (according to ISO 639-3 codes) of 58 different language families, written in 16 different scripts (according to ISO 15924 codes). Again, 1,000 lines of linguistic utterances are randomly sampled from the TeDDi sample to roughly match the number of sequences available also for protocuneiform periods, and the Aurignacian material. This subsample includes 33 languages and 14 different scripts. See SI Appendix, Fig. S7 and Table. 1 for some example sequences.

Preprocessing. All sequences are coded in UTF-8 characters. For a visualization of the preprocessing procedure see SI Appendix, Fig. S7. For natural languages (TeDDi sample) punctuation is removed, as well as characters which are generally not part of the respective writing system. A single "sign" here corresponds to a UTF-8 character (including white spaces)—compare also sign counts under #Tokens in SI Appendix, Fig. S7.

In the case of protocuneiform, annotations of the original transliterations in CDLI are removed. Also, the numerical notation to indicate repetitions of the same sign types are automatically expanded to yield strings of sign tokens, e.g. 3(N01) → N01 N01 N01. Note that N01 stands for a single indent with a stylus on the clay tablet (often representing the number one). Underscores are added between visual groups of signs. Note that a "sign" here corresponds to a string of characters in the Latin transliteration (e.g. N01, N14, UDU~a).

In the case of Aurignacian objects, sign sequences are cleaned by a) removing annotations (e.g. "!" indicating broken pieces), b) removing signs which are not visually discrete (marked by round brackets), and c) removing signs which are not in linear order (marked by square brackets). Underscores indicate that the respective sign strings are on different parts of an object. Signs are here represented by UTF-8 characters, e.g. "l:" straight line, "v:" notch, "x:" cross, etc. For the full typology of sign types, see www.signbase.org. In SI Appendix, Fig. S5, we give frequency distributions of sign types by cave site, material, and object type.

Statistical features. Given the preprocessed sequences of Aurignacian signs, protocuneiform writing, and modern writing, their "statistical fingerprint" is established, i.e. quantitative features which have been proposed to distinguish natural languages from other sequences (25–27, 29, 85). In particular, we harness the type-token ratio (TTR), the unigram entropy (H), the entropy rate (h), as well as the repetition rate of adjacent signs (r).

Let us assume a vocabulary of sign types $\mathcal{V} = \{t_1, t_2, \dots, t_m\}$, with $m \in \mathbb{N}$ being the finite size or cardinality of the set, i.e. $m = |\mathcal{V}|$. Sign types are unique, i.e. $t_1 \neq t_2 \dots \neq t_m$. A sequence is then a concatenation of sign tokens (t'_i) such that

$$\mathbf{s} = (t'_1, t'_2, \dots, t'_n), \quad [2]$$

where $n \in \mathbb{N}$ is the length of the sequence in tokens. Note that sign tokens are not unique, i.e. it is possible that $t'_1 = t'_2 \dots = t'_n$ for some type t_i . In the following, m and n represent the vocabulary sizes and lengths of particular sequences (i.e. m_S and n_S in a strict sense). The \mathbf{s} is dropped for simplicity.

Type-Token-Ratio (TTR). The type-token ratio of a given sequence \mathbf{s} is then defined as

$$\text{TTR}(\mathbf{s}) = \frac{|\mathcal{V}|}{\sum_{i=1}^n f_i} = \frac{m}{n}, \quad [3]$$

with f_i being the absolute token frequency count in the sequence \mathbf{s} for a given type t_i . In words, the number of different types in a sequence (m) is divided by the number of tokens (n). The TTR is a simple, well-known, all-purpose quantitative linguistics measure, which can be seen as a baseline. It reflects the diversity of sign types given the overall number of sign tokens. It is in the range $(0, 1]$, if we only consider nonempty sequences, i.e. $0 < m \leq n$. For example, in the Aurignacian sequence XX_vvvvvvvv_vvvv (see vhc0096 in SI Appendix, Fig. S7) the "vocabulary" is the set $\mathcal{V} = \{x, v, _\}$, and the token frequencies of sign types are "x:" 2, "v:" 12, "_:" 2. We thus get $\text{TTR} = 3/16 \sim 0.19$. Compare also Table. 1.

Unigram Entropy (H). Compared to TTR, the Shannon entropy of unigrams is a more nuanced measure of diversity, reflecting the probability distribution over the elements of the vocabulary (86, pp. 51–55). Let us assume a discrete random variable S which takes the vocabulary \mathcal{V} as its support, that is, each draw will result in a token representing a type $t_i \in \mathcal{V}$. The probabilities of types are represented by the probability mass function $p(t) = P(S = t)$. The entropy of S is then theoretically defined as (87, p. 14)

$$H(S) = - \sum_{t \in \mathcal{V}} p(t) \log_2 p(t) \quad [4]$$

In fact, it can be seen as an average over the information content of types which is given by

$$I(t) = - \log_2 p(t). \quad [5]$$

This takes values in the range $[0, \infty)$. For a type of maximal probability, i.e. $p(t_i) = 1$, we have $-\log_2(1) = 0$, and for a type of minimal probability, i.e. $p(t_i) = 0$, we have $-\log_2(0) = \infty$. The original rationale of Claude E. Shannon (88) is that likely events provide us with less information than unlikely events. The entropy is the average information content of a vocabulary of types, hence also falling theoretically in the range $[0, \infty)$. However, for a finite sequence \mathbf{s} , the maximal entropy is empirically upper-bounded by the case of uniform probability over types in \mathcal{V} , i.e. $p(t) = \frac{1}{m}$, yielding

$$\hat{H}(\mathbf{s}) \leq - \sum_{t \in \mathcal{V}} \frac{1}{m} \log_2 \frac{1}{m} \equiv \log_2(m). \quad [6]$$

Again, this property was deliberately chosen by Shannon (88) to reflect the fact that maximal uncertainty should be given in the case all events are equally likely to occur.

A crucial problem for entropic investigations of finite sequences is how to estimate the probabilities $p(t)$ of types. We here employ the so-called maximum likelihood (ML), or plug-in estimator

$$\hat{p}(t_i)^{\text{ML}} = \frac{f_i}{\sum_{i=1}^m f_i} = \frac{f_i}{n}, \quad [7]$$

where f_i is, as above, the absolute frequency of a type t_i in the sequence \mathbf{s} . The probability of a given type is hence seen as a relative frequency over the length of the sequence. Plugging Eq. 7 into Eq. 4 gives the maximum likelihood estimated unigram entropy of a sequence. We call this the unigram entropy (\hat{H}^{ML}). For example, for the preprocessed Mandarin Chinese sequence in SI Appendix, Fig. S7 (ID13125567) we have

$$\hat{H}^{\text{ML}}(\mathbf{s}) = -\left(\frac{2}{7} \log_2\left(\frac{2}{7}\right) + \frac{1}{7} \log_2\left(\frac{1}{7}\right) + \dots + \frac{1}{7} \log_2\left(\frac{1}{7}\right)\right) \sim 2.89 \text{ bits/sign.} \quad [8]$$

There is a range of further entropy estimators which have been developed to alleviate data sparsity problems by smoothing probabilities. However, it turns out that all of them are highly correlated for language data (86).

Table 1. Examples of sequences and feature values

Subcorpus	Sequence	#t'	#t	\hat{H}	\hat{h}	TTR	r
TeDDi (Basque)	Mila esker maitea	17	11	3.29	2.19	0.65	0
TeDDi (Chinese)	see SI Appendix, Fig. S7 ID13125567	7	6	2.52	1.69	0.86	0.17
TeDDi (Chinese strokes)	etao etao aieeeeatn eaaieeeeeto taeietso oodtshn etasee	54	10	2.89	1.99	0.19	0.17
Uruk III	N01 _ NAGA~a _ DU _ PAP~a	7	5	2.13	1.43	0.71	0
Uruk IV	N14 N14 _ N01 N01 N01 _ UDU~a	8	4	1.91	1.37	0.5	0.43
Uruk V	N01 N01 N01	3	1	0	0.43	0.33	1
Aurignacian	XX_vvvvvvv_vvvv	16	3	1.06	1.04	0.19	0.73

Selected sequences with number of sign tokens (t'), number of sign types (t), as well as type-token ratios (TTR), estimated unigram entropies (\hat{H}), entropy rates (\hat{h}), and repetition rates (r).

Entropy Rate (\hat{h}). While the TTR and the unigram entropy H only take into account the frequencies/probabilities of individual types— independent of their context—the concept of entropy rate can be employed to assess the predictability given that signs in a sequence might display systematic co-occurrence patterns.

Formally, instead of a single random variable S we consider a stochastic process $\mathbf{S} = \{S_1, S_2, \dots, S_n\}$ with vocabulary \mathcal{V} . This can be seen as a concatenation of random variables, rather than a single random variable. In theory, the entropy rate for a stochastic process is defined as (87, p. 74)

$$h(\mathbf{S}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(S_1, S_2, \dots, S_n), \tag{9}$$

With $H(S_1, S_2, \dots, S_n)$ representing the joint entropy of the random variables. Thus, the entropy rate is only defined in the limit of a sequence with infinite length. Given a finite sequence \mathbf{s} , Gao et al. (89) equation (6) propose an increasing-window estimator based on Lempel-Ziv compression (LZ78). This is defined as

$$\hat{h}(\mathbf{s}) = \frac{1}{n} \sum_{i=2}^n \frac{\log_2(i)}{L_i}, \tag{10}$$

where i is a given position in the sequence $\mathbf{s} = (s_1, s_2, \dots, s_n)$, and L_i is the length (+1) of the longest contiguous subsequence starting at position i which is also present in the so-called “prefix,” i.e. in the subsequence s_1^{i-1} . Note that the numerator $\log_2(i)$ gives the maximally possible entropy of the sequence up to position i , namely, when we have seen only unique types such that $i = m$ (remember Eq. 6). The denominator, on the other hand, is a representation of the repetitiveness of the sequence. Take the sequences $\mathbf{s}_1 = (\text{same_but_different})$ and $\mathbf{s}_2 = (\text{same_same})$ as an example. In position $i = 6$, for both we have $\log_2(6) \sim 2.58$ bits/sign in the numerator. However, for \mathbf{s}_1 we have $L_6 = 0 + 1 = 1$, while for \mathbf{s}_2 we have $L_6 = 4 + 1 = 5$ in the denominator. Overall the ratio is then $\frac{2.58}{1} \sim 2.58$ bits/sign for \mathbf{s}_1 , and $\frac{2.58}{5} \sim 0.52$ bits/sign for \mathbf{s}_2 . In other words, the repetitiveness inherent to \mathbf{s}_2 is penalized when estimating the entropy rate.

One problem with this estimator is that it converges relatively slowly for highly repetitive sequences. This has to do with the fact that as we approach positions toward the end of the sequence, the L_i s necessarily become shorter since the subsequences after position i , i.e. s_i^n , become shorter. This leads to an overestimation bias. See also simulations in Gao et al. (89, p. 91). We illustrate the stabilization properties for our specific dataset in [SI Appendix, Figs. S9–S12](#).

Repetition Rate (r). The repetition rate has been proposed by Richard Sproat (27) as a feature to distinguish writing from nonwriting. It measures the number of adjacent repetitions of signs in a sequence over the number of possible repetitions. Formally, it can be defined as

$$r^{\text{sproat}}(\mathbf{s}) = \frac{r^{\text{adj}}}{\sum_{i=1}^m (f_i - 1)}, \tag{11}$$

where r^{adj} in the numerator stands for the number of adjacent repetitions of sign types. More precisely, we have a repetition if for two tokens in the sequence

\mathbf{s} it holds that $t'_j = t'_{j+1} = t_j$. The denominator is a normalization given the possible number of repetitions in the sequence, where f_i is the token frequency of a type t_i and m is the overall number of types—as above. This normalization term is named R in the original publication (27). Note that for any given type the maximum possible number of adjacent repetitions is $f_i - 1$. In Sproat’s definition, this is summed up over all types to get R .

We here slightly modify this definition by changing the denominator to $n - 1$ instead, i.e. simply the length of the sequence \mathbf{s} minus one (abbreviated to l_{smo} here). We thus have

$$r^{\text{lsmo}}(\mathbf{s}) = \frac{r^{\text{adj}}}{n - 1} = \frac{r^{\text{adj}}}{(\sum_{i=1}^m f_i) - 1} \tag{12}$$

Both versions of the repetition rate r take values in the range $[0, 1]$. The main difference is that for sequences with many unique types r^{sproat} will give high values, while r^{lsmo} will give low values. For example, for the sequences (aaaaa) and (abcd), we have $r^{\text{sproat}} = 1$ in both cases, while we have $r^{\text{lsmo}} = 1$ and $r^{\text{lsmo}} = 0.25$ respectively.

We have implemented both versions of the repetition rate in the statistical software R (90). The function is called `rrate()`. We generally use the *lsmo* version in this study.

Further details on the convergence properties of the statistical features as well as randomization analyses are given in [SI Appendix](#).

Classification algorithms. Given the feature value estimations, for each sequence \mathbf{s} we get a four-dimensional vector (\vec{v}) of values. In other words, a sequence is here represented as a point in a four dimensional space, i.e.

$$\vec{x}_{\mathbf{s}} = \begin{bmatrix} x_1 = \hat{H}(\mathbf{s}) \\ x_2 = \hat{h}(\mathbf{s}) \\ x_3 = \text{TTR}(\mathbf{s}) \\ x_4 = r(\mathbf{s}) \end{bmatrix} \tag{13}$$

For example, for the Basque, the Mandarin Chinese, and the Aurignacian sequences in [Table. 1](#), we have

$$\vec{x}_{\text{Basque}} = \begin{bmatrix} 3.29 \\ 2.19 \\ 0.65 \\ 0 \end{bmatrix}, \vec{x}_{\text{Mandarin}} = \begin{bmatrix} 2.52 \\ 1.69 \\ 0.86 \\ 0.17 \end{bmatrix}, \vec{x}_{\text{Aur}} = \begin{bmatrix} 1.06 \\ 1.04 \\ 0.19 \\ 0.73 \end{bmatrix}. \tag{14}$$

A core question in this research is whether it is possible to clearly classify these sequences based on their feature values. This boils down to a standard classification task: Can we assign the correct labels, i.e. names of subcorpora (TeDDi, UrukIII, UrukIV, UrukV, Aurignacian), to the sequences based on their feature values? If yes, then the statistical fingerprint of sequences can indeed be used to identify the subcorpus they belong to.

Training and Test Sets. In total, we have 3078 sequences with feature value estimations and subcorpus labels. Note that we always test classification performance on pairs of subcorpora (i.e. Aurignacian vs. UrukV, Aurignacian vs. UrukIII, Aurignacian vs. UrukIV, and Aurignacian vs. TeDDi). These pairwise

Table 2. Number of sequences for subcorpus pairs overall, training and test sets

Subcorpus1	Subcorpus2	Overall	Training	Test
Aurignacian	TeDDi	1,253	850	403
Aurignacian	UrukIII	1,068	728	340
Aurignacian	UrukIV	1,072	732	340
Aurignacian	UrukV	324	216	108

subsamples are then split into training and test sets by the ratio 67% to 33%. An overview of the subsample structure is given in Table 2.

These training and test sets of sequences are used as input to classification algorithms. We use K-Nearest-Neighbors (KNN) as a baseline classification algorithm, and Multilayer Perceptrons (MLP), also known as deep feedforward neural networks, as state-of-the-art classification algorithms.

With regard to neural network architectures, i.e. hidden layer depth and size, we randomly sample 100 architectures with a maximum depth of four layers and a maximum size of five hidden units per layer. More complex architectures are unlikely to converge with our dataset. An example of a complex architecture which converges is given in SI Appendix, Fig. S13.

In summary, for each binary classification between subcorpora we have theoretically up to 100 results—one for each architecture. In practice, there are fewer results, since not all architectures converge. The neural net analyses were run on a 16 GB NVIDIA RTX A4000 GPU. Further mathematical details on the classification algorithms are given in SI Appendix.

Performance Statistics. The accuracy, precision, recall, and F1 scores are reported for each classification run—corresponding to one value of k for a KNN, or one MLP architecture trained and tested on the sequences.

We here focus on the accuracy as this comes with a straightforward baseline and a standard statistical test for significant deviation from the baseline. The accuracy is given as the number of correct labels assigned to sequences of the test set (successes), over the total number of labels assigned (successes plus failures). A worked example of how to calculate the test statistic is given in SI Appendix.

For each k in KNN classification, and for each model architecture in MLP classification, we get a P -value reflecting whether the success rate of the given classification algorithm is significantly better than the baseline. Since we run many statistical tests, we additionally correct the P -values by the Bonferroni method, i.e. multiplying the P -value by the number of tests while keeping the α -level constant.

If the sequences have different statistical properties, then they should be distinguishable by the classification algorithms, and there should be many P -values below the α -level. We provide the percentage of significant P -values alongside the distributions of accuracy values in SI Appendix, Fig. S4.

Statistical models for Aurignacian data. Besides the classification analyses, we also provide regression models for the Aurignacian data to further contextualize

Table 3. Multiple regression model results

Model specification	DF	r. DF	F	P	AIC
Intercept	1	212	NA	NA	198
Intercept + volume	2	207	23	2e-06	175
Intercept + volume + date	3	203	14	2e-06	171
Intercept + volume + date + object_type	8	198	10	2e-10	145
Intercept + volume + date + object_type + site_name	10	196	8	5e-10	147
Intercept + volume + date + object_type + preservation	10	196	8.5	1e-10	144
Intercept + volume + date + object_type + material	11	195	7.2	1e-09	148
Intercept + volume*preservation + date + object_type	12	194	7	5e-10	146
Intercept + volume*object_type + date	13	193	6.5	1e-09	147
Intercept + volume + date*object_type	13	193	6.3	2e-09	149

The model specification gives the predictor variables included in the model. DF: degrees of freedom. r. DF: residual degrees of freedom (number of data points minus parameters). F: F-statistic. P : P -value for the F-test. AIC: Akaike Information Criterion. The best model in terms of AIC is highlighted in bold.

the usage of sign sequences by ancient humans of the Swabian Jura. In particular, we are interested which meta-data categories per object predict higher or lower information density of the sequences on it.

Response Variable. We use the entropy rate as response variable, since this feature reflects both the diversity of the sign types used on an object (similar to TTR and unigram entropy), but also their sequential structure (similar to the repetition rate). However, the R code is implemented such that the statistical feature can be chosen in a single line of code, and all the analyses can be rerun without changing the rest of the code. Furthermore, note that the entropy rate is correlated with the number of sign tokens and the number of sign types per sequence. However, using these as response variables will lead to violation of regression assumptions (e.g. normality of residuals).

Predictor Variables. As predictor variables we consider the object type (factor variable: "anthropomorph figurine," "zoomorph figurine," etc.), material (factor variable: "antler," "ivory," etc.), cave site (factor variable: "Hohlenstein-Stadel," "Hohle Fels," etc.), preservation (factor variable: "almost complete," "fragmented," "complete"), volume in cm^3 (continuous variable), and maximum estimated age of the object (approx. 40,000 BP to 30,000 BP). See also SI Appendix, Fig. S14.

Model Formulation. Given the response variable Y_{hrate} we first define a baseline model with a single parameter β_0 , i.e., the intercept or average of Y such that $\beta_0 = \bar{Y}$:

$$Y_{\text{hrate}} = \beta_0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2). \quad [15]$$

The errors ϵ (or residuals) are assumed to be normally distributed around this mean. We then build models stepwise by adding theoretically relevant predictors one at a time, following the advice in Baayen (91), and comparing the respective model to the former model by means of AIC—with lower AIC indicating a preferable model (Table 3). For example, with the continuous predictor "object volume" (X_1) added we have

$$Y_{\text{hrate}} = \beta_0 + \beta_1 X_1 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad [16]$$

where β_1 is the coefficient which reflects the effect that the predictor variable has on the entropy rate. The best model arrived at in terms of AIC includes volume and date as continuous predictors, as well as object type (six levels) and preservation (three levels) as categorical predictors. Given the treatment dummy coding for covariance analyses in R (91, p. 339) this requires overall ten parameters (nine coefficients plus intercept) to be estimated:

$$Y_{\text{hrate}} = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_9 X_9 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad [17]$$

where β_0 is the group mean of Y for the data points with the two default factor levels, and X_0 is a variable always taking the value one, β_1 and X_1 as well as β_2 and X_2 might represent the effects of the continuous variables (volume and date) on Y , and the other coefficients β_3, \dots, β_9 would then represent the effects of the binary dummy coded variables representing the levels of the categorical predictors X_3, \dots, X_9 . For this best model, all model assumptions (normality of residuals, homoscedasticity, etc.) are checked in *SI Appendix, Fig. S15*.

Data, Materials, and Software Availability. The analyses are implemented in the statistical software R (90). The data of SignBase Version 2.0 is available at Zenodo: <https://doi.org/10.5281/zenodo.18401937> (92). Data and R code of this study are available at Zenodo: <https://doi.org/10.5281/zenodo.17673743> (93). This gives the permanent version of the github repository: <https://github.com/christianbentz/PaleoSigns>.

ACKNOWLEDGMENTS. Thanks to Mattis List for access to the high-performance laboratory computer of the Chair for Multilingual Computational Linguistics, University of Passau, and to Kellen Parker van Dam for providing support with running the code. Thanks to Saetbyul Lee for her consistent support with database programming. Marieluise Hahn's help with data curation for SignBase is also

gratefully acknowledged. We thank Nicholas J. Conard (University of Tübingen), Stefanie Kölbl (Urgeschichtliches Museum Blaubeuren), Fabian Haack and Erwin Keefer (Landesmuseum Württemberg), Kurt Wehrberger (Museum Ulm), and Ernst Seidel (Museum der Universität Tübingen) for their support in analyzing the artifacts of the Swabian Jura and for the permission to depict the objects. Thanks to the *Staatliche Museen zu Berlin, Vorderasiatisches Museum/Olaf M. Teßmer* for providing high-resolution pictures for protocuneiform tablets. Thanks to the Co-Evolution of Language, Interaction and Architecture interest group organized by Antje Wilton and Kenan Hochuli, and the audiences of the Workshop on "Icons in transition" co-organized by Marion Benz and Barbara Helwing at the World Neolithic Congress in Urfa 2024, as well as the Workshop "Genesis of Writing" at the Centre for the Study of Manuscript Cultures (University of Hamburg) organized by Michael Friedrich for their feedback and discussions. Finally, we would like to thank Kenan Hochuli for coining the term "Evolutionary Semiotics." This work is funded by the European Union (European Research Council, The Evolution of Visual Information Encoding, 101117111). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

- J. C. Joordens *et al.*, Homo erectus at Trinil on Java used shells for tool production and engraving. *Nature* **518**, 228–231 (2015).
- D. Leder *et al.*, A 51,000-year-old engraved bone reveals Neanderthals' capacity for symbolic behaviour. *Nat. Ecol. Evol.* **5**, 1273–1282 (2021).
- F. d'Errico *et al.*, From number sense to number symbols. An archaeological perspective. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **373**, 20160518 (2018).
- C. S. Henshilwood, F. d'Errico, I. Watts, Engraved ochres from the Middle Stone Age levels at Blombos Cave, South Africa. *J. Hum. Evol.* **57**, 27–47 (2009).
- E. Dutkiewicz, G. Russo, S. Lee, C. Bentz, Signbase, a collection of geometric signs on mobile objects in the Paleolithic. *Sci. Data* **7**, 364 (2020).
- J. J. Hublin *et al.*, Initial Upper Palaeolithic Homo sapiens from Bacho Kiro Cave, Bulgaria. *Nature* **581**, 299–302 (2020).
- D. Myopotamitaki *et al.*, Homo sapiens reached the higher latitudes of Europe by 45,000 years ago. *Nature* **626**, 341–346 (2024).
- M. Stoneking, J. Krause, Learning about human population history from ancient and modern genomes. *Nat. Rev. Genet.* **12**, 603–614 (2011).
- J. A. Haws *et al.*, The early Aurignacian dispersal of modern humans into westernmost Eurasia. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25414–25422 (2020).
- L. Slimak *et al.*, Modern human incursion into Neanderthal territories 54,000 years ago at Mandrin, France. *Sci. Adv.* **8**, eabj9496 (2022).
- L. Slimak, The three waves: Rethinking the structure of the first Upper Paleolithic in Western Eurasia. *PLoS One* **18**, e0277444 (2023).
- B. Delluc, G. Delluc, *Les manifestations graphiques Aurignaciennes sur support rocheux des environs des Ezies (Dordogne)*, Gallia Préhistoire (Éditions du CNRS, Paris, 1978), vol. 21.
- M. Chollot-Varagnac, *Les origines du graphisme symbolique. Essai d'analyse des écritures primitives en Préhistoire* (Éditions de la Fondation Singer-Polignac, Paris, 1980).
- R. Bourillon *et al.*, A new Aurignacian engraving from Abri Blanchard, France: Implications for understanding Aurignacian graphic expression in Western and Central Europe. *Quat. Int.* **491**, 46–64 (2018).
- E. Dutkiewicz, N. Conard, "The symbolic language of the Swabian Aurignacian as reflected in the material culture from Vogelherd Cave (South-West Germany)" in *L'art au quotidien. Objets ornés du Paléolithique supérieur. Actes du colloque international Les Eyzies-de-Tayac, 16-20 Juin 2014*, J. Cleyet-Merlie, J. Geneste, E. Man-Estier, Eds. (SAMRA, 2014), pp. 149–164.
- E. Dutkiewicz, S. Wolf, N. J. Conard, Early symbolism in the Ach and the Lone valleys of southwestern Germany. *Quat. Int.* **491**, 30–45 (2018).
- E. Dutkiewicz, *Zeichen: Markierungen, Muster und Symbole im Schwäbischen Aurignacien* (Kerns Verlag, Tübingen, 2021).
- M. Otte, *Le Paléolithique supérieur ancien en Belgique* (Musées Royaux d'art et d'histoire, Brussels, 1979).
- M. Lejeune, "Le Trou Magrite et l'art mobilier Aurignacien en Belgique: Synthèse et problèmes. Das Trou Magrite und die Aurignacienzeitliche Kleinkunst in Belgien: Synthèse und Probleme" in *Les chemins de l'art Aurignacien en Europe. Das Aurignacien und die Anfänge der Kunst in Europa. Colloque International, Internationale Fachtagung, Aurignac, 16-18 Septembre 2005*, H. Floss, N. Rouquerol, Eds. (Editions Musée-Forum Aurignac, Aurignac, 2007), pp. 131–144.
- A. Majkić, S. Evans, V. Stepanchuk, A. Tsvelykh, F. d'Errico, A decorated raven bone from the Zaskalnaya VI (Kolosovskaya) Neanderthal site, Crimea. *PLoS One* **12**, e0173435 (2017).
- A. Majkić, F. d'Errico, S. Milošević, D. Mihailović, V. Dimitrijević, Sequential incisions on a cave bear bone from the Middle Paleolithic of Pešturina Cave, Serbia. *J. Archaeol. Method Theory* **25**, 69–116 (2018).
- R. G. Bednarik, Concept-mediated marking in the Lower Palaeolithic. *Curr. Anthropol.* **36**, 605–634 (1995).
- R. G. Bednarik, The Lower Paleolithic engravings of Bilzingsleben, Germany. *Encyclopedia* **4**, 695–708 (2024).
- A. Marshack, Lunar notation on Upper Paleolithic remains. *Science* **146**, 743–745 (1964).
- R. P. N. Rao *et al.*, Entropic evidence for linguistic structure in the Indus script. *Science* **324**, 1165 (2009).
- R. P. Rao *et al.*, Entropy, the Indus script, and language: A reply to R. Sprout. *Comput. Linguist.* **36**, 795–805 (2010).
- R. Sprout, A statistical comparison of written language and nonlinguistic symbol systems. *Language* **90**, 457–481 (2014).
- M. Corazza, F. Tamburini, M. Valério, S. Ferrara, Unsupervised deep learning supports reclassification of Bronze age cyprriot writing system. *PLoS One* **17**, e0269544 (2022).
- C. Bentz, "The Zipfian challenge: Learning the statistical fingerprint of natural languages" in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, J. Jiang, D. Reitter, S. Deng, Eds. (Association for Comput. Linguist., Singapore, 2023), pp. 27–37.
- L. L. P. Herskind, F. Riede, A computational linguistic methodology for assessing semiotic structure in prehistoric art and the meaning of southern Scandinavian Mesolithic ornamentation. *J. Archaeol. Sci.* **165**, 105969 (2024).
- S. Ferrara *et al.*, The invention of writing on Rapa Nui (Easter Island). New radiocarbon dates on the Rongorongo script. *Sci. Rep.* **14**, 2794 (2024).
- L. A. Courtenay, F. d'Errico, R. Núñez, D. E. Blasi, Identifying potential palaeolithic artificial memory systems via spatial statistics: Implications for the origin of quantification. *Archaeol. Anthropol. Sci.* **17**, 1–20 (2025).
- I. Aron *et al.*, Whale song shows language-like statistical structure. *Science* **387**, 649–653 (2025).
- M. Youngblood, Language-like efficiency in whale communication. *Sci. Adv.* **11**, eads6014 (2025).
- S. Semple, R. Ferrer-i Cancho, M. L. Gustison, Linguistic laws in biology. *Trends Ecol. Evol.* **37**, 53–66 (2022).
- R. Heesen, C. Hobaier, R. Ferrer-i Cancho, S. Semple, Linguistic laws in chimpanzee gestural communication. *Proc. R. Soc. B* **286**, 20182900 (2019).
- F. d'Errico, A new model and its implications for the origin of writing: The La Marche antler revisited. *Camb. Archaeol. J.* **5**, 163–206 (1995).
- A. Leroi-Gourhan, Le symbolisme des grands signes dans l'art pariétal paléolithique. *Bull. Soc. Préhist. Fr.* **55**, 384–398 (1958).
- A. Leroi-Gourhan, *Gesture and Speech* (MIT Press, Cambridge, Massachusetts, 1993 [1964]).
- G. Von Petzinger, A. Nowell, A place in time: Situating Chauvet within the long chronology of symbolic behavioral development. *J. Hum. Evol.* **74**, 37–54 (2014).
- G. Von Petzinger, *The First Signs: Unlocking the Mysteries of the World's Oldest Symbols* (Simon and Schuster, New York, 2017).
- B. Bacon *et al.*, An Upper Palaeolithic proto-writing system and phenological calendar. *Camb. Archaeol. J.* **33**, 371–389 (2023).
- K. A. Overmann, The role of materiality in numerical cognition. *Quat. Int.* **405**, 42–51 (2016).
- E. Mellet *et al.*, Neuroimaging supports the representational nature of the earliest human engravings. *R. Soc. Open Sci.* **6**, 190086 (2019).
- K. Tylén *et al.*, The evolution of early symbolic behavior in Homo sapiens. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 4578–4584 (2020).
- F. N. Wullf, S. K. Kristensen, I. Wisner, K. Tylén, "Eyes on the past: Visual exploration of Upper Palaeolithic cave art" in *Proceedings of the 46th Annual Conference of the Cognitive Science Society* (2024), vol. 46, pp. 2457–2464.
- M. Pagnotta *et al.*, The evolution of symbolic artefacts: How function shapes form. *Evol. Hum. Behav.* **45**, 261–267 (2024).
- S. Garrod, N. Fay, J. Lee, J. Oberlander, T. MacLeod, Foundations of representation: Where might graphical symbol systems come from? *Cogn. Sci.* **31**, 961–987 (2007).
- B. Galantucci, S. Garrod, Experimental semiotics: A review. *Front. Hum. Neurosci.* **5**, 1–15 (2011).
- J. Nölle, B. Galantucci, "Experimental semiotics: Past, present, and future" in *The Routledge Handbook of Semiosis and the brain*, A. M. Garcia, A. Ibanez, Eds. (Routledge, New York/Abingdon, 2022), pp. 66–81.
- O. Morin, Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy. *Cogn. Sci.* **42**, 664–677 (2018).
- O. Morin, P. Kelly, J. Winters, Writing, graphic codes, and asynchronous communication. *Topics Cogn. Sci.* **12**, 727–743 (2020).
- T. Higham *et al.*, Testing models for the beginnings of the Aurignacian and the advent of figurative art and music: The radiocarbon chronology of Geißenklösterle. *J. Hum. Evol.* **62**, 664–676 (2012).

54. N. J. Conard, V. Rots, Rope making in the Aurignacian of Central Europe more than 35,000 years ago. *Sci. Adv.* **10**, eadh5217 (2024).
55. N. J. Conard, M. Malina, S. C. Münzel, New flutes document the earliest musical tradition in southwestern Germany. *Nature* **460**, 737–740 (2009).
56. S. Wolf, *Schmuckstücke: die Elfenbeinbearbeitung im Schwäbischen Aurignacien*. Tübingen Monographien zur Urgeschichte (Kerns Verlag, Tübingen, 2015).
57. N. J. Conard, A female figurine from the basal Aurignacian of Hohle Fels Cave in southwestern Germany. *Nature* **459**, 248–252 (2009).
58. N. Ebinger-Rist, S. Wolf, K. Wehrberger, C. J. Kind, L'homme-lion d'Hohlenstein-Stadel. *L'Anthropologie* **122**, 415–436 (2018).
59. S. Moran, C. Bentz, X. Gutierrez-Vasques, O. Pelloni, T. Samardzic, "TeDDi sample: Text data diversity sample for language comparison and multilingual NLP" in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari et al., Eds. (European Language Resources Association, Marseille, France, 2022), pp. 1150–1158.
60. C. S. Peirce, *Prolegomena to a science of reasoning: Phaneroscopy, semeiotic, logic* (Peter Lang GmbH, Frankfurt am Main, 2016[1906]).
61. F. J. Thackeray, D. Morris, A 'symbolically wounded' engraved zebra from Wonderwerk cave, and linguistic evidence. *Digging Stick* **23**, 18 (2006).
62. P. Kelly, J. Lei, H. J. Bibiko, L. Barker, Amsd: The Australian message stick database. *PLoS One* **19**, e0299712 (2024).
63. P. Kelly, Australian message sticks: Old questions, new directions. *J. Mater. Cult.* **25**, 133–152 (2020).
64. G. Sauvet, S. Sauvet, A. Wlodarczyk, Essai de sémiologie préhistorique (pour une théorie des premiers signes graphiques de l'homme). *Bull. Soc. Préhist. Fr. Études Trav.* **74**, 545–558 (1977).
65. M. García-Bustos, O. Rivero, G. Sauvet, P. G. Bustos, Discussion: "An Upper Palaeolithic proto-writing system and phenological calendar" by Bennett Bacon et al. (2023). *J. Paleolithic Archaeol.* **6**, 1–13 (2023).
66. F. J. Thackeray, The bonne saison hypothesis and prehistoric European art. A preliminary study of mammoths represented at Rouffignac cave. *Digging Stick* **40**, 9–10 (2023).
67. A. Nowell, P. Bahn, J. L. Le Quellec, Evaluating the evidence for lunar calendars in Upper Palaeolithic parietal art. *Camb. Archaeol. J.* **35**, 38–55 (2024).
68. C. Woods, "Visible language: The earliest writing systems" in *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*, C. Woods, Ed. (The Oriental Institute of the University of Chicago, Chicago, Illinois, 2010), pp. 15–28.
69. K. Pozdniakov, G. Segerer, Similar place avoidance: A statistical universal. *Linguist. Typol.* **11**, 307–348 (2007).
70. A. Nevins, "Haplological dissimilation at distinct stages of exponence" in *The Morphology and Phonology of Exponence*, J. Trommer, Ed. (Oxford University Press, Oxford, 2012), pp. 84–116.
71. C. A. Cathcart, Multiple evolutionary pressures shape identical consonant avoidance in the world's languages. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2316677121 (2024).
72. I. J. Gelb, *A Study of Writing* (The University of Chicago Press, Chicago & London, 1952).
73. R. Sproat, *Symbols: An Evolutionary History from the Stone Age to the Future* (Springer Nature, Cham, Switzerland, 2023).
74. D. Schmandt-Besserat, *How Writing Came About* (University of Texas Press, Austin, 1992).
75. R. K. Englund, "Texts from the Late Uruk period" in *Mesopotamien. Annäherungen 1: Späturuk-Zeit und Frühdynastische Zeit*, P. Attinger, M. Wäfler, Eds. (Vandenhoeck & Ruprecht, Freiburg, 1998), pp. 15–233.
76. E. Dutkiewicz, *Die Kosmologie der Altsteinzeit in Kosmos - Vom Umgang mit der Welt zwischen Ausdruck und Ordnung* (Heidelberg University Publishing, Heidelberg, 2023), pp. 29–64.
77. K. A. Overmann, Material scaffolds in numbers and time. *Camb. Archaeol. J.* **23**, 19–39 (2013).
78. K. A. Overmann, Early writing: A cognitive archaeological perspective on literacy and numeracy. *Vis. Lang.* **56**, 9–44 (2022).
79. J. F. Cantlon, S. T. Piantadosi, Uniquely human intelligence arose from expanded information capacity. *Nat. Rev. Psychol.* **3**, 275–293 (2024).
80. R. Ferrer-i Cancho, Optimization models of natural communication. *J. Quant. Linguist.* **25**, 207–237 (2018).
81. CDLI contributors, Home (2024). <https://cdli.mpiwg-berlin.mpg.de/> (Accessed 15 May 2024).
82. H. J. Nissen, Uruk and I. *Cuneif. Digit. Libr. J.* **1**, 1–18 (2024).
83. H. J. Nissen, The archaic texts from Uruk. *World Archaeol.* **17**, 317–334 (1986).
84. R. K. Englund, "Proto-cuneiform account-books and journals" in *Creating Economic Order: Record-Keeping, Standardization and the Development of Accounting in the Ancient Near East*, M. Hudson, C. Wunsch, Eds. (CDL Press, Bethesda, Maryland, USA, 2004), pp. 23–46.
85. C. Bentz, D. Alikaniotis, M. Cysouw, R. Ferrer-i Cancho, The entropy of words - learnability and expressivity across more than 1000 languages. *Entropy* **19**, 275 (2017).
86. C. Bentz, *Adaptive Languages: An Information-Theoretic Account of Linguistic Diversity* (Walter de Gruyter GmbH & Co KG, Berlin/Boston, 2018), vol. 316.
87. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New Jersey, 2006).
88. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
89. Y. Gao, I. Kontoyiannis, E. Bienenstock, Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy* **10**, 71–99 (2008).
90. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2024).
91. R. H. Baayen, "Multivariate statistics" in *Research Methods in Linguistics*, R. J. Podesva, S. Devyani, Eds. (Cambridge University Press, Cambridge, 2013), pp. 337–372.
92. E. Dutkiewicz, H. Marieluise, S. Lee, Saetbyul, C. Bentz, SignBase Version 2.0. Zenodo. <https://doi.org/10.5281/zenodo.18401937>. Deposited 28 January 2026.
93. C. Bentz, PaleoSigns v1.1. Zenodo. <https://doi.org/10.5281/zenodo.17673743>. Deposited 21 November 2025.